Draft copy. Final version to appear in: E. Salas, C. Bowers, & E. Edens (Eds.), Improving teamwork in organizations: Applications of resource management training (pp. 165-188). Mahwah, NJ: Erlbaum.

Evaluating Resource Management Training

Robert W. Holt, Deborah A. Boehm-Davis, J. Matthew Beaubien
George Mason University

Resource management is a critical component of effective group performance in a number of domains, including aviation, medicine, and the military. Although a fair amount of research has been devoted to the development of resource management training programs (Helmreich & Foushee, 1993; Wiener, Kanki, & Helmreich 1993), much less effort has been devoted to their evaluation. The evaluation of a training program is important for a number of reasons, not the least of which is to determine whether the organization's investment pays off in terms of demonstrable performance improvements. In many domains, however, changes in performance are difficult to measure because of uncontrollable factors that exist within the larger organizational context.

This chapter will outline the steps required to evaluate the effectiveness of a resource management training program, and highlight the various practical and theoretical issues that arise during this process. We will first cover general requirements for defining, implementing, and evaluating resource management training. Then we will illustrate these principles by applying them to Crew Resource Management (CRM) in the aviation domain. While this chapter will emphasize the application of statistical techniques and research design, page constraints limit our discussion of these topics. Interested readers should refer to more comprehensive expositions provided by Campbell and Stanley (1963), Cook and Campbell (1979), Howell (1997), and Pedhazur and Pedhazur Schmelkin (1991).

DEVELOPING A RESOURCE MANAGEMENT EVALUATION PLAN

Principles of Evaluation

While several different approaches are available for evaluating the effectiveness of a resource management training program (Joint Committee, 1994; Goldstein, 1993; Guttentag & Struening, 1975), certain principles remain invariant. For example, the primary objective is to determine (1) if resource management training makes a noticeable difference in the dependent variables, and (2) the magnitude of the training program's effect.

At a minimum, training should make a difference that is noticeable. A noticeable difference has two components. First, it is a difference that statistical methods determine to be non-chance (above a background level of noise due to measurement error). Second, the difference should have practical value to the organization. If it is determined that training made a noticeable difference, then the size of the training effect should be estimated so that cost-benefit analyses can be performed. If multiple training programs have been developed, the data can be used to assess the relative effectiveness of the different training methods.

When evaluating a training program, it is critical to collect measures of performance at the appropriate time. If performance is evaluated before training has "sunk in," a training effect may not be observed (Kraiger et al., 1993). Similarly, evaluating performance after too long an interval may contaminate the data with uncontrolled intervening events that obscure the effects of training. Thus, the right time interval must be chosen to accurately evaluate training effects. If a valid theory of performance is available for the training domain, the time interval can be based on this theory. Alternatively, if the right time interval is unknown, evaluation should be repeated over a reasonable period of time to check for both immediate and delayed training effects.

It is also important to determine where to look for changes in performance. For example, Kirkpatrick (1976) suggests that training effectiveness can be manifest at several levels of analysis: the individual, the team or crew, and the organization. A majority of the resource management literature focuses exclusively on the *immediate* transfer of trained material at the individual or team level. This is not unreasonable, as individual/team behaviors are most directly under the control of trainees. However, *long-term* aggregate performance data, for example at the department or organizational level, are also important to the organization. Unfortunately, performance data, unlike measures of behavior, are frequently beyond the control of the individuals or team (Campbell, 1990). For example, an aircrew may manage a crisis situation perfectly, yet factors beyond their control, such as faulty equipment can nonetheless lead to a disaster.

Therefore, it is important to remember that any measured effect can have multiple causes. Although training is one such cause, a systematic evaluation should attempt to rule out as many plausible alternatives as possible, so that the training program can be isolated as the primary source of the observed differences (Campbell & Stanley, 1963; Cook & Campbell, 1979). For these reasons, the effects of resource management training should be evaluated in a systematic, step-by-step fashion. This requires developing a list of *targeted* changes in knowledge, skills, and/or attitudes that are expected to occur after training, and investigating them in a *systematic* fashion (Kraiger, Ford, & Salas, 1993).

Selecting an Evaluation Design

All evaluations of resource management training programs rely on some form of comparison. The simplest type of evaluation involves comparing groups with different degrees or methods of training to one another using the same set of criteria. Still another form of comparison is to compare pre-training performance to post-training performance. The various approaches differ on the type of comparison emphasized, and on the amount of control over confounds. Regardless of which approach is chosen, the goal is to develop the fairest and least confounded comparison of the effects of training (Campbell & Stanley, 1963; Cook & Campbell, 1976).

Evaluation approaches range along a continuum from extremely controlled studies modeled on the experimental method, with people randomly assigned to separate trained and untrained groups, to relatively uncontrolled field studies, in which the training is done *in vivo* and the effects are measured in the natural environment. There are costs and benefits associated with each approach. The sections that follow will highlight these trade-offs.

Experimental Designs

A traditional experimental design requires the ability to randomly assign persons to trained and untrained groups (or different levels/types of training). The trained group is then compared to the untrained group on each possible criterion variable. This is the most precise evaluation of training effectiveness, but probably the least practical, as most organizations will usually want to train all job incumbents. One variation of the traditional experiment is a "waiting list" control group. In this variation, all people ultimately receive the training, but the people designated to receive the training first vs. last are *randomly* determined. In the window of time where the first group(s) are trained and the last group(s) are not, the effects of training can be measured on what are essentially randomly-assigned trained and untrained groups.

Quasi-Experimental Designs

If naturally occurring groups are available but cannot be randomly assigned, a quasi-experiment can be performed in which one group is trained and the other group is not. As in a traditional experiment, both groups are evaluated for the effects of resource management training. The major disadvantage of this design, however, is that the groups may not be equivalent on other relevant variables such as ability, experience, and so forth.

In commercial aviation, the naturally occurring groups are fleets, and fleets typically differ in the average age and experience of the pilots therein. Therefore, the possibility exists that some characteristics unique to the trained group may interact with the training to produce the measured effects. This makes it essential to measure possible confounds (e.g., differences in experience across fleets), and assess their effects on the evaluation criteria, such as via hierarchical regression or analysis of covariance.

Pre/Post Evaluation

If everyone must receive training at the same time, evaluation studies can be set up to address changes in the trainees' performance. For example, after resource management training, trainees should have higher levels of efficiency and productivity while simultaneously having lower levels of errors and other undesirable outcomes. This is one of the easiest methods of evaluation, and at the very minimum, some form of pre/post design should be used to evaluate the effects of training.

Unfortunately, this evaluation method is also one of the weakest because it is subject to many confounds such as contextual effects and maturation. If these confounds occur between the pre-training and post-training measurements, they can artificially cause the observed changes in performance. Therefore, the pre-training measurement should be taken early enough to be unaffected by the knowledge of or anticipation of training, but not so early that the baseline performance could change a great deal prior to training.

Time-Series Evaluation

A time-series design extends the time where performance is measured before and after training. Extending these intervals of measurement provides the advantage of being able to rule out potential confounds, such as a general increase performance due to maturation. However, it does so at the cost of additional measurements.

When making multiple measurements, the effect of the measurement process itself must be considered. For example, if supervisors are simply rating subordinates on naturally observed performance, subjects may not react negatively to the measurement process (although effects of making multiple assessments on the part of the supervisor should still be considered). However, if subordinates are put in an specially-designed evaluation scenario for each measurement, then practice effects, learning of test-relevant knowledge and skills, and changes in performance motivation may very well occur. Any situation in which the subordinate is strongly aware of the testing and evaluation process is open to these types of confounds.

DEVELOPING MEASURES OF RESOURCE MANAGEMENT PERFORMANCE

Once the research design is selected, measures must be developed that address the constructs of interest. Accurate performance assessment requires several critical steps: defining the construct, developing appropriate measurement instruments, and objectively confirming the psychometric properties of these instruments. While these steps are highly interdependent, they will be discussed separately for clarity of exposition.

To accurately assess resource management, it must first be defined. Without a specific operational definition, appropriate assessments of resource management cannot be developed. If the construct is multi-dimensional, then *multiple* measures need to be developed. Once developed, these measures must be evaluated for acceptable levels of statistical sensitivity, reliability, and validity. After the quality of these measures has been established, they may be confidently used to obtain a full and accurate evaluation of the resource management training program.

Defining Resource Management

Resource management is potentially difficult to define and measure because it is complex, multi-dimensional, and process-oriented (see Lauber, 1984 for more information). Given this complexity, it may be necessary to create several operational definitions, one for each of the various resource management dimensions and processes.

An operational definition is a precise, focused definition that is used for a specific purpose such as evaluation. Any operational definition must be complete and specific enough to clearly imply appropriate measurement strategies and techniques. As a general rule, the operational definition should specify (1) the core knowledge, skills, and behaviors required for effective resource management, and (2) relevant situational factors that describe the context in which performance is measured.

Developing Appropriate Measures

Effective resource management should affect both task- and relationship-oriented aspects of performance (Borman & Motowidlo, 1993). The performance changes may occur at the individual, team, and/or organizational level (Kirkpatrick, 1976), but their form and inter-relationships may vary across levels (Chan, 1998). However, practical limitations generally require the evaluation process to focus on a selected *subset* of these possible effects. At a minimum, this subset should include process and outcome measures at both the individual and team level.

Performance changes at the individual, team, or organization level may occur at different time frames. Kirkpatrick (1976) proposed a model which suggests that training results are manifested at multiple stages: initial reactions to the training program, changes in knowledge and behavior during the training, transfer of trained behaviors to the workplace, and changes in organizational effectiveness. According to Kirkpatrick's model, each stage is a necessary but insufficient precursor to the following stages. Despite previous criticisms and caveats (Alliger & Janak, 1989; Alliger, Tannenbaum, Bennett, Traver, & Shotland, 1997; Goodman, Lerch, & Mukhopadhyay, 1994), this model provides a useful framework for considering the effect of training interventions at different organizational levels. In general, individual effects of training appear first, followed by team, and then organizational changes. Therefore, the appropriate time to measure individual, team, and organizational effects may vary considerably.

Unfortunately, measurement of resource management performance is more difficult than measuring the output of an assembly-line worker. When a physical object is being produced, productivity can be indexed in terms of output quality or quantity. In contrast, the evaluation of process variables such as resource management requires evaluating the interaction of a team within a complex system. For example, evaluating resource management in aviation crews depends on the interaction between the Captain and the First Officer as well as their interactions with flight attendants, air traffic control personnel, and

the physical aircraft systems (Boehm-Davis, Holt, & Seamster, this volume). Therefore, it may be desirable to measure each construct via a number of different methods. The principle of converging operations (Campbell & Fiske, 1959) suggests that if different measurement methods provide the same result, confidence in that result is increased. Whenever possible, multiple measures of resource management performance should be included in the evaluation process.

At the same time, it is also wise to measure more than just one possible effect of training (Kraiger, Ford, & Salas, 1993). For example, relevant outcomes of resource management training at the individual level may include attitudes toward resource management, declarative knowledge of resource management procedures, and changes in trainee's knowledge structures (Schvaneveldt, 1990). Relevant outcomes forthe team may include increased task and social cohesion, a perception of more collective competence, an increase in shared knowledge structures, and better group interaction processes. In aviation, relevant crew outcomes would include improved communication, coordination, situation awareness, planning and decision-making. Relevant outcomes at the organizational level would depend on the domain. In the aviation domain, relevant outcomes may include on-time performance, decreased fuel consumption, fewer incidents, and decreased insurance costs.

Measuring Performance

Due to the complexity of resource management performance, the evaluation method of choice is often a performance *rating* regarding the quality of resource management behavior at the individual or crew level. This evaluation should be guided by appropriate tools and materials that help the evaluator make an accurate assessment. For example, carefully-designed rater training programs and evaluation worksheets developed according to the principles of human factors have the potential to reduce the rater's cognitive workload. This may simplify the evaluation process, and give more reliable results (see Boehm-Davis, Holt, & Seamster, this volume). Other materials required for evaluation will depend on the evaluation context.

The context for evaluation can be either job performance in a normal context or performance measured in a special evaluation context. One common method is to have evaluators make an overall assessment of typical performance, which is often done annually . This particular evaluation has the advantage of reflecting the person's resource management in diverse job-related situations over an extended period of time. Nevertheless, there are disadvantages when using this evaluation technique. These include incomplete or distorted recall for relevant events, recency bias, memory priming caused by the phrasing of evaluation questions, and the influence of pre-existing knowledge about the individual being evaluated (DeNisi, Cafferty & Meglino, 1984). Other evaluation problems depend on the number of persons evaluated. If each evaluator rates only a few individuals, their evaluations may be poor due to limited practice with the rating system and exposure to a limited

range of performance. At the same time, if each evaluator assesses multiple persons, carry-over or contrast effects may adversely influence individual performance ratings.

Special evaluation contexts can be designed to avoid or minimize these errors, but may have the disadvantage that performance in the special context is at a maximal rather than a typical level, and thus may not generalize to the job (Dubois, Sackett, Zedeck, & Fogli, 1993; Sackett, Zedeck, & Fogli, 1988). In the aviation domain, the work sample of a normal flight is typically combined with realistic simulations of normal working conditions to increase generalization and obtain more typical levels of resource management behaviors. Special evaluation requires the preparation of extra materials, including the work sample itself and guides/scripts to standardize evaluators' behavior during the assessment. Furthermore, evaluators must be appropriately trained in the use and administration of these materials (Prince, Oser, Salas, & Woodruff, 1993).

Measuring Knowledge

One option for evaluating resource management is to evaluate the components that contribute to performance, such as the information that individuals have acquired as a result of the training program. Training may change two types of knowledge: declarative knowledge and procedural knowledge. Declarative knowledge refers to the static information about a domain that is represented in memory. It can be thought of as the definitions for constructs in the domain, and rules for when this knowledge can (or should) be applied. Procedural knowledge, on the other hand, typically refers to rules regarding the *execution* of specific behaviors (Anderson, 1985). Although procedural knowledge is based in part on declarative knowledge, it is considered to be a "higher order" form of knowledge, because it involves the integration of multiple sources of information, as well as the automation of specific behaviors.

For example, before one can turn an aircraft by coordinating aileron and rudder movements, one must have the appropriate foundation of declarative knowledge about adverse yaw caused by moving the ailerons. Because procedural and declarative knowledge are manifest in different forms, they must be assessed differently. Typically, *elements* of declarative knowledge are assessed via paper-and-pencil measures, while the *organization* of declarative knowledge is assessed via techniques such as Pathfinder (Schvaneveldt, 1990). Procedural knowledge, on the other hand, is typically assessed with some form of work sample test (Kraiger, Ford, & Salas, 1993).

Effectiveness Criteria

Another issue to consider when developing appropriate measures is the different ways training can affect performance. The goal of training may be to change the mean (average) level of performance or to change the distribution (variability) of performance.

Traditionally, training is evaluated in terms of mean differences. For example, the mean performance of trained crews is often compared to that of untrained crews.

However, some researchers (Alliger & Katzman, 1997) argue that certain training interventions can influence both the mean and/or variability of performance data. For example, group consensus training or instructor calibration training is often used to decrease the random variability in people's response patterns, while simultaneously having little or no effect on mean ratings. Conversely, training may attempt to increase the variance of ratings. For example, training in creativity may seek to increase the variability of ideas generated by a group. Therefore, it is essential that researchers avoid the temptation to assess training performance solely in terms of mean change.

The specific outcomes of training should be guided by an overall theory of resource management in the domain of interest. This theory of performance should, in turn, be used to develop a systematic measurement plan (Kraiger, Ford, & Salas, 1993) which specifies which type and level of performance to be expected, the time at which this performance is expected to occur, and the appropriate measurement strategy for each facet of performance.

Multifaceted Approaches and Multiple Constituencies

In recent years, a number of researchers have heeded Kraiger, Ford, & Salas's (1993) call for a multifaceted approach to the evaluation of training programs (Leedom & Simon, 1995; Salas, Fowlkes, Stout, Milanovich, & Prince, in press; Stout, Salas, & Fowlkes, 1997; Stout, Salas, & Kraiger, 1997). In general, these studies have included a variety of individual-level (e.g., reactions to training, declarative knowledge, knowledge organization) and group-level (crew processes, crew outcomes) criteria as indices of the effectiveness of CRM training programs. Unfortunately, even these well-designed and well-intentioned studies attest to the difficulties of performing systematic training evaluation in organizational contexts. For example, several studies were limited by psychometrically deficient measures of declarative knowledge, small sample sizes, or the measurement of immediate, maximal performance to the exclusion of long-term, typical performance.

Although these groundbreaking efforts were more complete and multifaceted than previous evaluations, their weaknesses illustrate two basic principles that still jeopardize the usefulness of a training evaluation. First, no matter how many criterion variables are measured, the information that they provide is only as good as the measurement instrument. For example, a given study may measure both reactions to training and declarative knowledge. However, to the extent that the measures of declarative knowledge are psychometrically deficient (e.g., the lack of item difficulty results in "ceiling" effects), they provide little additional information regarding the effectiveness of the training program (Crocker & Algina, 1986).

Second, virtually every training program is going to have *some* effect on immediate performance, but these could be transitory effects. Commercial air carriers invest tens of millions of dollars every year with the implicit understanding that

training programs will result in performance increases that carry over to typical performance in line operations over the long term with the ultimate criteria being increased safety and efficiency in line operations. Therefore, training professionals must conduct studies that assess the *long-term effects* of CRM training programs. For example, if the effect of a training program wears off after one month of line performance, it would probably not be considered an effective program from the airline' s perspective. Different constituencies such as the researcher, the carrier, the union, the Federal Aviation Administration, and the general public may have different criteria for success, and that these success criteria are often at odds with one another (Austin, Klimoski, & Hunt, 1996). Therefore, carriers and researchers alike need to be more considerate of the needs of these other constituencies. We believe that long-term measures of CRM training program effectiveness will address at least some of these needs.

Ensuring the Quality of Measurement

The third step in the evaluation process is to objectively confirm the psychometric quality of these assessment measures. Since evaluations are performed by individuals, the quality of evaluations is decreased by inaccuracy, subjectivity, or personal biases on the part of the evaluator. Objectively confirming the quality of measurement instruments involves three basic facets. A good measure of resource management must be *sensitive* enough to discriminate good from poor resource management, *reliable* enough to consistently provide the same estimate of resource management, and *valid* enough to ensure that the measure involves only resource management rather than other extraneous factors. We will cover each facet of measurement in turn.

Measurement Sensitivity

Sensitivity refers to the extent to which a measure can detect changes in the construct being assessed. Specifically, a sensitive measure of resource management should show higher scores when resource management is above average, and lower scores when resource management is below average. While extreme examples of good or bad performance are usually easy to detect, sensitivity must also be established for subtle differences in resource management behaviors, such as *marginally* safe vs. unsafe performance.

Sensitivity is influenced by the granularity of the measurement instrument. More specifically, the evaluation scale must be sufficiently fine-grained to capture important differences in the quality of resource management that is observed, yet still be accurately used by the evaluator. For example, a dichotomous “satisfactory vs. unsatisfactory” scale might be accurately used by evaluators, but would not be sensitive to varying degrees or good or bad resource management. Conversely, a 100-point scale might be extremely fine-grained, but evaluators may not be able to use it accurately. A compromise for

measurements based on human evaluations is often a five- or seven-point scale with meaningful definitions assigned to each scale point (Likert, 1936).

To objectively index the sensitivity of measurement, it is necessary to compare the judgements made by evaluators to pre-established levels of resource management. One method for indexing the sensitivity of evaluation is to have evaluators rate "test" cases of varying levels of resource management proficiency (as determined by subject-matter experts). For example, average evaluator ratings for "good" test cases ought to be higher than ratings for "average" test cases, which in turn should be higher than ratings for "poor" test cases. One way to index sensitivity for each evaluator is to use Hays' (1988) omega-squared index for strength of effect (Williams, Holt, & Boehm-Davis, 1997; Holt, Johnson, and Goldsmith, 1997). This index reflects how different an evaluator's ratings are for different categories of test cases and has a range from 0.0 (no discrimination among levels) to 1.0 (perfect discrimination among levels).

Measurement Reliability

Informally, reliability can be defined as the consistency or stability of measurement. Formally, reliability is defined as the lack of random error in the measurement instrument (Nunnally, 1967). While different traditional methods of estimating reliability have been developed, we will only cover two: test-retest reliability and internal consistency reliability (see Nunnally, 1967 or Pedhazur & Pedhazur Schmelkin, 1991 for more information). Because each method makes different assumptions about the main source of error in measurement, each has its own advantages and disadvantages.

Test-retest reliability is used to assess the stability of measurement over time. One method of assessing this form of reliability consists of having evaluators assess the same set of performances at two different times and correlating these two sets of evaluations. The calculation is based on the Pearson product-moment correlation, and results in an index *r* that reflects reliability. In this case, a value of *r* near 0.0 indicates a lack of test-retest reliability, whereas values near 1.0 indicate near-perfect test-retest reliability. However, test-retest reliability assumes that the only important source of random error is spontaneous changes over time. Unfortunately, systematic evaluator differences are common in evaluating resource management in the aviation domain (Williams, Holt, & Boehm-Davis, 1997). To the extent these differences are stable over time, the test-retest reliability is inflated. Therefore, although simple to execute, the test-retest reliability method only addresses one potential source of error, and may be positively biased.

Internal consistency reliability refers to the internal coherence of a set of items which are all measuring the same thing (Nunnally, 1967). For evaluation of resource management, this type of reliability requires a set of multiple items all reflecting resource management. If resource management has distinct components, each distinct component must have its own set of multiple items. The intercorrelations among items in a set are summarized into a Coefficient Alpha index which ranges

from 0.0 (no internal consistency reliability) to 1.0 (perfect internal consistency reliability). Several factors influence coefficient alpha, such as the number of items included in the scale (Cortina, 1996), as well as *systematic* judgment errors made by evaluators (e.g., halo rating errors). To the extent that these systematic errors occur across items, internal consistency reliability will be inflated.

When used in isolation, both test-retest and internal consistency reliability estimates can provide misleading results. To check and correct such rater errors, we have developed an alternative approach for training and checking evaluator reliability that uses multiple statistical indexes for evaluating rater performance and giving training feedback. This multi-component approach was labeled Inter-Rater Reliability (IRR) training (Holt et al., 1997).

During the IRR process, each evaluator's ratings of the test cases are compared to the group's judgments by using four indexes, each of which provides information on one aspect of reliability. In addition, an index of the sensitivity of judgment is also included if SME's have evaluated the test cases. First, the overall distribution of each evaluator's ratings is compared to the group's distribution to ascertain its level of *congruency*. Low congruency suggests the evaluator gives a different mix of ratings on the scale compared to the group. Second, *systematic differences* of harsher or more lenient grading among the evaluators are identified. Third, the inter-rater correlation is calculated to see if the raters shift in a consistent manner up and down in their ratings across evaluated items (*consistency*). Finally, if the test cases have been externally scored by subject-matter experts, raters can also be assessed regarding the *sensitivity* of their evaluations. Rater-specific estimates of congruency, systematic differences, consistency, and sensitivity results are provided to each individual, while the aggregate results for all raters are provided to the group for discussion.

The group of raters is also provided with information concerning their level of group *agreement* on each item (James, Demaree, & Wolf, 1993). This feedback is critical because every item with low agreement should be discussed until reasonable group consensus is reached. In summary, the IRR method compares each rater to the group using indexes that give the rater information about the congruency, systematic differences, consistency and sensitivity of his or her evaluations. The information from these indexes and group agreement is then used to train and improve subsequent ratings (Williams, Holt, & Boehm-Davis, 1997).

Measurement validity

Validity refers to the extent to which a measure really measures its intended construct (Nunnally, 1967; Landy 1986). More specifically, validity is the proportion of variance in a measure that reflects real variation in the measured construct . From a resource management perspective, validity refers to the amount of variability in evaluator ratings that

accurately reflects real differences in the resource management performance of the persons being evaluated. Assessing validity requires checking measurement items, the measurement process, and the results of the measurement process.

The items used for evaluating resource management should be checked for face and content validity (Nunnally 1967). Face validity refers to the judgment of a group of experts that the items are plausibly measuring the desired construct. Such judgments are easy and convenient, but unfortunately they are also somewhat subjective. A more objective item analysis will often indicate that items designed by experts to measure a given construct do not, in fact, predict that construct. Face validity is, therefore, easy to establish but only weak evidence that the construct in question is being assessed.

Content validity first requires a careful specification of the domain of all possible relevant items. Content validity can then be demonstrated by showing that the evaluation items are a fair, unbiased, and representative sample of items from this larger domain. Techniques for specifying relevant content items for training programs have been developed (Lawshe, 1975). However, because resource management typically requires an individual or team to interact in a complex system, the set of possible items is very large and ill-defined. Therefore, the specification of the domain of all possible relevant items for this type of complex domain may be difficult or impossible.

The validity of measurement generally is established by empirically examining the relationship to other measures that should be related to the construct. Two basic principles apply. The first principle is convergent and discriminant validity (Campbell & Fiske, 1959). In convergent validity, measures which ought to be related to a construct should converge or correlate with the proposed measure. For resource management, measures which ought to positively relate to it, such as measures of teamwork-KSAs (Stevens & Campion ,1994), ought to positively correlate with the resource management measure. A valid measure of a construct should show the expected relationships with plausible criteria (criterion validity), and predict the expected outcomes of changing resource management (predictive validity). In divergent validity, measures which ought to be independent or distinct from resource management should diverge or not correlate with the proposed measure. For example, if resource management can be done equally well by men and women, then gender should not correlate with resource management measures. Divergent validity is particularly important if potential confounds like popularity or appearance could influence a measure of resource management effectiveness; they must be shown not to do so.

The second principle is network validity (Pedhazur & Pedhazur Schmelkin, 1991). For network validity, the nomological network of constructs that should be theoretically associated with the construct is empirically assessed to determine if it demonstrates the expected pattern of relationships. For example, a valid measure of resource management ought to show a plausible set of relationships with antecedents, correlates, and consequences that one would expect for resource management. If the expected network of relationships is generally found, network validity is established.

EXAMPLE EVALUATION PROGRAM DEVELOPMENT

We recently worked with a regional air carrier to develop and evaluate a resource management training program for pilots. This training program focused on improved crew briefings and communication during normal operations, as well as problem diagnosis, situation assessment, and planning and decision making during abnormal or emergency operations. This program was unique in that the resource management principles were translated into step-by-step operational procedures. Further, these procedures were formally required as part of Standard Operating Procedure (SOP) for one fleet and added to the operating manuals and handbooks for that particular aircraft.

Selecting the Level at which Resource Management would be Evaluated

We evaluated the effectiveness of the training program by measuring performance at both the individual and crew levels. Clearly, the performance of individual pilots is important. First, individual pilots must be qualified to continue to legally operate an aircraft. Second, the performance of an individual can directly affect the performance of his or her team or crew. Third, some issues such as the effects of ability on performance, were more sensibly addressed by comparing the assessed ability of individuals to their performance (Boehm-Davis, Holt, & Hansberger, 1997).

Although individual performance is important, commercial aircraft are always operated by crews. The performance of a team or crew may be quite distinct from the performance of individual of team members, especially among highly-complex, interdependent tasks (Steiner, 1972). Further, evidence from aviation accidents and safety reports suggests that a lack of coordination among crewmembers has been the cause of a substantial portion of problems on the flight deck (National Transportation Safety Board, 1994). As a result, this project also focused on crew-level performance.

Developing the Evaluation Plan

Selecting an Evaluation Design

This particular carrier was composed primarily of two fleets. The research team decided to provide the resource management training program to one fleet, while the other fleet continued to use existing procedures and management techniques. In this quasi-experimental design, the fleet with extra training and new procedures acted as the experimental group while the fleet with normal training and procedures acted as the control group. One focus of the evaluation design was to compare pilots and crews in the two fleets.

In order to allow for gradual learning on the part of the pilots of the new procedures and processes, we also incorporated aspects of a time-series design. Specifically, we collected pilot and crew performance measures over a three-year period. During the first year, the pilots had additional resource management training but the new procedures had not been formally implemented. This was our baseline performance year. In the second year, the new procedures were formally

implemented and required as SOP for that fleet. Performance measured in that year would reflect the immediate impact of the resource management training and SOP changes. The third year was the final follow-up assessment that would either confirm or disconfirm long-term effects of the training, including a gradual acceptance and accommodation to the new methods of cockpit interaction and coordination. In addition, during the final year of evaluation, three auxiliary measures of resource training were developed. These additional methods allowed converging measures of the effects of this training with different samples of evaluators and performance situations.

Developing Measures of Resource Management Performance

Once the evaluation design had been selected, the next steps were to develop an operational definition of resource management, develop appropriate measures given that operational definition, and to ensure the quality of the measures that were developed.

Defining Resource Management

For this project, effective crew resource management (CRM) was defined for two qualitatively distinct contexts: normal operations and abnormal/emergency situations. For normal situations, effective CRM was defined as the effective communication and coordination of crewmembers before, during, and after flying a typical flight. The operational definition of normal performance included quality of briefings and other communication, quality of workload management and avoiding overload, maintaining situation awareness of the aircraft and external traffic and weather situation, and preserving effective coordination on checklists, flows and other sequential tasks during the flight.

For abnormal/emergency situations, effective CRM was defined as effective workload management and communication while performing normal flight tasks plus problem diagnosis, situation assessment, planning, and monitoring of plan execution. The operational definition of abnormal/emergency CRM was quite extensive and included, for example, the establishment of explicit "bottom lines" and "backup plans" during the planning task, plus clearly communicating these plan components to other crew members.

Developing Appropriate Measures

In carrying out this project, we developed a variety of measures to capture both individual and crew level performance. Further, we realized that these metrics would be applied by a number of different evaluators (pilot instructor/evaluators). Thus, we felt that it was also important to develop a structured method for collecting assessments of pilot and crew performance.

Measuring performance. A structured evaluation process was designed to achieve systematic and reliable observations and ratings of performance. The multi-year evaluations consisted of a Line Operational Evaluation (LOE) and

Line Checks. The LOE evaluation, conducted during the pilots' annual evaluation for flight certification, consisted of a work-sample performance evaluation in which the crews performed a typical flight scenario in a full-motion simulator. The evaluator followed an LOE script to consistently introduce specific problems and distracting conditions into the flight. In this way, crew reactions to abnormal and emergency situations could be assessed in a standardized manner. The evaluation forms emphasized specific crew reactions for these events, including both technical and CRM performance items and related skills. The basis for the evaluation forms was the specification of a set of observable behaviors. These observable behaviors were carefully identified by subject matter experts as being central to successful performance on a specific event set. These behaviors and skills provided a point of focus for the instructor/evaluators during the observation and evaluation of the LOE and during the crew debriefing after the LOE.

The Line Check assessed pilot and crew performance during normal flight operations. Typically, instructor/evaluators would board a routine flight without prior announcement and evaluate the crew on a spectrum of technical and CRM items. For this carrier, crew performance ratings, both technical and CRM, were based on a standardized four-point scale covering the full range of possible crew performance from Unsatisfactory Performance (observed crew behavior does not meet minimal requirements) to Above Standard Performance (observed crew behavior is markedly better than the Standard Performance).

To provide converging measurement of crew performance, we designed two auxiliary performance measures. First, the cadre of instructor/evaluators who had evaluated pilots from both the experimental and control fleets completed a detailed performance questionnaire regarding the *relative* performance of pilots from both fleets during upgrade or transition training from the control fleet to the experimental fleet. Second, a separate cadre of five evaluators assessed pilots from both fleets during normal flights using a direct observation form. This cadre was completely different from the carrier's Line Check evaluators or FAA evaluators, and the assessments were strictly voluntary. By using different sets of evaluators and different evaluation formats (e.g., the LOE and Line Check), these data provide converging information regarding the hypothesized performance differences between the two fleets.

To ensure a broader measurement of possible training effects besides performance, we also used a pilot survey to measure knowledge and attitudes as suggested by Kraiger, Ford, and Salas (1993). Knowledge acquired by individual pilots from the training program was measured by a survey of all carrier pilots in the final year of the project. Knowledge was only measured post-training because the training introduced completely new procedures developed for this project which pilots could not have known about previously. Therefore, the focus of the knowledge evaluation was on the extent to which individual pilots were able to describe the new set of procedures and the appropriate context for enacting each procedure. The focus of

attitude measurement concerned attitudes toward CRM in general and more specifically towards the trained resource management procedures. The survey also measured how often pilots performed the new procedures and briefings, and the perceived effects of the new procedures and briefings.

Focus on mean changes. The major focus in this project was on mean differences between the two fleets. That is, we were interested in demonstrating that the crews in the trained fleet would perform at a higher level on measured crew resource management skills than crews that had not received the training. Mean differences were the focus because narrowing the range or variability of performance would not have been a useful outcome.

For attitudes, we compared the mean attitudes of pilots in the two fleets to one another, as well as to a neutral baseline. For assessing knowledge, we measured the relative extent of relevant knowledge and tested whether the trained pilots could answer knowledge questions at an above-chance level (representing *some* knowledge). Similarly, the pilots' perceptions of the frequency of performance and effects of the new procedures were analyzed.

Ensuring the Quality of Measurement

Sensitivity

Instructor/evaluators (I/Es) were presented with videotapes showing different levels of resource management behavior, derived from simulation sessions conducted by the airline. The I/Es were asked to rate the level of resource management behavior exhibited on the videotapes using a four-point scale, ranging from unsatisfactory (1) through FAA minimal requirements (2), company standard (3), and above company standard (4). Each level of this scale had a unique well-anchored qualitative meaning for the raters. The segments of behavior portrayed on the videotapes were selected to represent the range of possible resource management behavior, with a focus on behaviors rated in the central portion of the scale (levels 2 and 3). Subject-matter experts established the exact level of performance for each segment. Sensitivity was indexed by analyzing the differences in each rater's evaluations for performance segments at different levels.

Reliability

Reliability was assessed on a regular basis (approximately every 6 months) using the multi-dimensional IRR procedures developed for this project (Williams, Holt, & Boehm-Davis, 1997). This process relies on a group of raters (instructor/evaluator pilots) using normative information for standardizing inter-rater reliability. All raters individually evaluated a videotape of typical crew performance on the LOE, and these evaluations were statistically compared. The relative amount of congruency of judgment distributions, systematic harsh or lenient judgments, inter-rater consistency, and agreement were assessed at each session. Each rater received same-day feedback about his or her evaluation performance relative to the other raters. Finally, each single item with agreement below a corporate standard was discussed intensively by the group to

isolate and solve causes of rater variability. The focus for this part of the training was the reduction of random variability on each item. Information from these discussions was also used to modify the content of the evaluation scenario, re-write evaluation items for improved clarity, formally codify explicit grading standards for certain items, and to modify or clarify carrier policies and procedures (SOP).

After this training, the performance evaluations conducted by these raters were accumulated into a database. After sufficient data were collected, the items that were designed to measure the same aspect of performance were assessed by an internal-consistency reliability metric (coefficient alpha). These estimates served as a final check on the reliability of the performance data.

Validity

The major focus for assessing validity was the internal structural validity of the assessment process. The evaluation data were analyzed by path analysis to verify that the process of evaluation was in fact performed in the correct manner. The process of evaluation from detailed behavioral observations to judgments of performance components to overall evaluations of performance was used to construct an anticipated structure of relationships among the performance measures. The expected path structure of relationships was found, which supported measurement validity.

Analysis and Interpretation of Evaluation Results

The LOE, Line Check, and auxiliary measures were all analyzed for the hypothesized fleet performance differences. Evidence from the LOE and the direct cockpit observations were crew-level assessments, and these results were examined for mean differences in crew performance. The evaluation of individual pilots was emphasized in the Line Check evaluations, the instructor/evaluator survey, and in a survey of individual pilots. Across these measures, both individual and crew levels of performance could be assessed, which were the targeted levels of change for this study.

Crew performance. On the LOE, several specific items concerning crew resource management behavior were graded with exactly the same grading standards for both fleets. For most of these items, the trained and untrained fleets were significantly different in the expected direction. The conclusion was that the resource management training had the desired effects for the work-sample evaluation.

The second crew-level evaluation was direct observations of cockpit interaction on regular line flights. These observations were carried out by a separate cadre of pilots who rode in the cockpit and watched the crew under voluntary, non-jeopardy conditions. Specific briefing content and other aspects performance relevant to the training were evaluated by these observers. These direct observations of cockpit interaction showed that crews from the trained fleet were significantly

superior on the majority of these items. On the remaining items, the trained fleet still had a higher mean, but the observed difference was not statistically significant.

Individual pilot performance. The first measure of performance at the pilot level was the instructor/evaluator survey. This survey involved a comparison of pilots from trained and untrained fleets who were transitioning aircraft or upgrading from First Officer to Captain. Instructor/evaluators who had experience with both sets of pilots gave comparative ratings for average individual pilot performance. These ratings indicated that in comparison to the untrained fleet, pilots from the trained fleet were significantly better in communication, workload management, and planning and decision-making.

The second measure of the effects of training on individual pilots was the pilot survey. The survey included pilots trained in specific resource management procedures and pilots without this training. Compared to appropriate baselines, trained pilots had acquired a significant amount of knowledge about the resource management procedures, had positive attitudes toward CRM and resource management procedures, frequently performed the trained procedures on routine flights, and strongly indicated that the procedures increased their effectiveness.

Convergent results for performance measurement and confirmatory results for attitudes and knowledge gives more confidence in the final evaluation of the effectiveness of this type of resource management training. Multiple evaluations at both the individual pilot level as well as the crew level help rule out various confounds or alternative explanations for the results. For example, positive effects of training were reported by instructor/evaluators, by an independent cadre of observer pilots, and by the pilots themselves. Each of these groups has different potential sources of bias, and the convergence of results reassures us that the positive effects are not simply the result of biased evaluators.

GUIDELINES FOR DEVELOPING EVALUATIONS OF RESOURCE MANAGEMENT TRAINING PROGRAMS

In developing a plan to evaluate a resource management training program, we recommend following the steps outlined in this chapter. These include: selecting a level at which to measure resource management, selecting a research design through which to evaluate the selected level of resource management behavior, and developing measurement instruments that can accurately assess resource management behaviors.

Specifically, Table 1 provides an overview of the steps needed to establish and implement an evaluation of a resource management training program. Each step has a set of critical issues that should be resolved for the best possible outcome.

Table 1.
Steps for Developing an Evaluation of a Resource

Management Training Program

1. Select the level at which resource management will be evaluated

2. Develop the evaluation plan
    - Select an evaluation design
    - Determine appropriate time interval for measuring change

3. Develop measures of resource management performance
    - Operationally define resource management
    - Develop appropriate measures
        - Measure knowledge, attitude, performance, etc.
        - Develop converging measures where possible
        - Decide to focus on mean changes or variability
    - Ensure the quality of measurement
        - Asses sensitivity
        - Assess reliability
        - Assess validity

4. Analyze and interpret the evaluation results

5. Use information to modify training system, personnel selection, etc.

We learned a number of important lessons at each step of the resource management evaluation process:

Guideline 1: An overall framework or theory about the type of performance measured must guide evaluation. Well-developed theory is critical for specifying the levels for the expected effects, operationally defining resource management, and for specifying the other measures necessary to establish construct validity.

Guideline 2: The level of evaluation of resource management is often determined by the context. In commercial aviation, the most important levels of evaluation were the individual pilots and the flight crews.

Guideline 3: Multifaceted evaluations of performance are preferred. The effects of resource management training should be examined for a broad range of possible changes. At a minimum, changes in knowledge, attitude, and behavior should be assessed.

Guideline 4: For measuring key effects like performance, multiple converging lines of evaluation evidence provide stronger support for the effects of resource management training. Creatively consider the various ways that the expected effects could be exhibited by individuals, teams, or organizational units, and measure them accordingly.

Guideline 5: A long-term multi-measure evaluation plan is necessary to detect delayed effects of training that may not be immediately apparent. Depending on the type of training the multiple observations may be collected over weeks, months, or years (as in the case study).

Guideline 6: Highly controlled evaluation is desirable. Nevertheless, the selected evaluation design should be a workable compromise between the desire for experimental control and the reality of the training and evaluation setting.

Guideline 7: Control groups are necessary. Having a control (untrained) group helps avoid many confounds that would otherwise hamper single-group evaluations.

Guideline 8: Repeated training in the evaluation process is necessary to maintain calibration of the raters for complex behavior domains such as resource management. Further, calibration must be continually re-checked for statistical levels of sensitivity, reliability, and validity.

Guideline 9: Evaluation of resource management is an iterative process. Ongoing evaluation may cycle back from Step 5 in Table 1 to an earlier step in the process. In our research, results from the LOE evaluations in years 1 and 2 helped change the LOE evaluation format to provide more precise, comparative evaluations in year 3.

Guideline 10: Careful evaluation of resource management will result in a bonus of new knowledge about performance appraisal, the training program, and relevant individual and team processes. More specifically, our research uncovered new information about pilots, crews, and the organization.

Guideline 11: Careful choices must be made at each step of the evaluation process. Each choice involves trade-offs between the desire for the best possible evaluation of resource management and the constraints of time, personnel, and other critical resources.

Acknowledgments

This research was supported by the Office of the Chief Scientific and Technical Advisor for Human Factors (AAR-100) at the Federal Aviation Administration through Grant 94-G-034 to George Mason University. We thank Dr. Eleana Edens at AAR-100 and Dr. Thomas Longridge at AFS-230 for their continuing support of this work.

References

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. Personnel Psychology, 42(2), 331-342.

Alliger, G. M., & Katzman, S. (1997). When training affects variability: Beyond the assessment of mean differences in training evaluation. In J. K. Ford & Associates (Eds.), Improving training effectiveness in work organizations (pp. 223-246). Mahwah, NJ: Earlbaum.

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. Personnel Psychology, 50(2), 341-358.

Anderson, J. R. (1985). Cognitive psychology and its implications. New York: W. H. Freeman and Company.

Austin, J. T., Klimoski, R. J., & Hunt., S. T. (1996). Dillematics in public sector assessment: A framework for developing and evaluating selection systems. Human Performance, 93(3), 177-198.

Boehm-Davis, D. A., Holt, R. W., & Seamster, T. L. (this volume). Airline experiences with resource management programs. In E. Salas, C. Bowers & E. Edens (Ed.), Applying resource management in organizations: A guide for training professionals. Mahwah, NJ: Earlbaum.

Boehm-Davis, D. A., Holt, R. W., & Hansberger, J. (1997). Pilot abilities and performance. Proceedings of the Ninth International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University Press.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), Personnel Selection in Organizations (pp. 71-98). San Francisco, CA: Jossey-Bass.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81-105.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Boston, MA: Houghton Mifflin.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., pp. 687-732). Palo Alto, CA: Consulting Psychologists Press.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. Journal of Applied Psychology, 83(2), 234-246.

Cook, T. D., & Campbell, D. T. (1976). Quasi-experimentation: Design & analysis issues for field settings. Boston, MA: Houghton Mifflin.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78(1), 98-104.

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, & Winston.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

DeNisi, A. S., Cafferty T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: a model and research propositions. Organizational Behavior and Human Decision processes, 33, 360-396.

Dubois, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximal performance criteria: Definitional issues, prediction, and white-black differences. Journal of Applied Psychology, 78, 205-211.

Goldstein, I. L. (1993). Training in organizations: Needs assessment, development, and evaluation (3rd edition). Pacific Grove, CA: Brooks/Cole.

Goodman, P. S., Lerch, F. J., & Mukhopadhyay, T. (1994). Individual and organizational productivity: Linkages and processes. In D. H. Harris (Ed.), Organizational linkages: Understanding the productivity paradox (pp. 54-80). Washington DC: National Academy Press.

Guttentag, M., & Struening, E.L. (1975). Handbook of Evaluation Research (Volume 2). Beverly Hills, CA: Sage.

Hays, W. L. (1988). Statistics (4th edition). Chicago, IL: Holt, Rinehart and Winston.

Helmreich, R. L., & Foushee, H. C. (1993). Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), Cockpit resource management (pp. 3-45). San Francisco, CA: Academic Press.

Holt, R. W., Johnson, P. J., & Goldsmith, T. E. (1997). Application of psychometrics to the calibration of air carrier evaluators. Proceedings of the Ninth International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University Press.

Howell, D. C. (1997). Statistical methods for psychology (4th edition). Boston, MA: Duxbury Press.

James, L. R., Demaree, R. G., Wolf, G. (1993). r-sub(wg): An assessment of within-group interrater agreement. Journal of Applied Psychology, 78(2), 306-309.

Joint Committee on Standards for Education Evaluation. (1994). The Program Evaluation Standards (2nd edition). Thousand Oaks, CA: Sage.

Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), Training and development handbook: A guide to human resource development (2nd edition). New York: McGraw-Hill.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based and affective theories of learning to new methods of training evaluation. Journal of Applied Psychology, 78(2), 311-328.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. American Psychologist, 41(11), 1183-1192.

Lauber, J. K. (1984). Resource management in the cockpit. Air Line Pilot, 53, 20-23.

Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel Psychology, 28, 563-575.

Leedom, D. K., & Simon, R. (1995). Improving team coordination: A case for behavior-based training. Military Psychology, 7, 109-122.

National Transportation Safety Board. (1994). A review of flightcrew-involved, major accidents of U.S. air carriers, 1978 through 1990. (Safety Study NTSB / SS-94 / 01, Notation 6241). Washington, DC: Author.

Nunnally, J.C. (1967). Psychometric Theory. New York, NY: McGraw-Hill.

Pedhazur, E. J., & Pedhazur Schmelkin, L. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Earlbaum.

Prince, C., Oser, R., Salas, E., & Woodruff, W. (1993). Increasing hits and reducing misses in CRM/LOS scenarios: Guidelines for simulator scenario development. International Journal of Aviation Psychology, 3(1), 69-82

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximal performance. Journal of Applied Psychology, 73(3), 482-486.

Salas, E., Fowlkes, J. E., Stout, R. J., Milanovich, D. M., Prince, C. (in press). Does CRM training improve teamwork skills in the cockpit?: Two evaluation studies. Human Factors.

Schvaneveldt, R. W., (Ed). (1990). Pathfinder associative networks: Studies in knowledge organizations. Norwood, NJ: Ablex.

Steiner, I. D. (1972). Group process and productivity. New York: Academic Press.

Stevens, M. J., & Campion, M. A. (1994). The knowledge, skill, and ability requirements for teamwork: Implications for human resource management. Journal of Management, 20(2), 503-530

Stout, R. J., Salas, E., & Fowlkes, J. E. (1997). Enhancing teamwork in complex environments through team training. Group Dynamics: Theory, Research, and Practice, 1(2), 169-182.

Stout, R. J., Salas, E., & Kraiger, K. (1997). The role of trainee knowledge structures in aviation team environments. International Journal of Aviation Psychology, 7(3), 235-250.

Wiener, E. L., Kanki, B. G. , & Helmreich, R. L. (1993), Cockpit resource management. San Francisco, CA: Academic Press.

Williams, D. M., Holt, R. W., & Boehm-Davis, D. A. (1997) Training for inter-rater reliability: baselines and benchmarks. Proceedings of the Ninth International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University Press.